Identification de gènes diagnostic chez les Rhizobium leguminosarum à l'aide de la grille informatique sous l'environnement WISDOM

Sébastien Guizard, Matthieu Matthieu.Reichstadt@inrae.Fr Reichstadt, Xavier Baiily

Identification de gènes diagnostic chez les Rhizobium leguminosarum à l'aide de la grille informatique sous l'environnement WISDOM

S. Guizard (1), M. Reichstadt (2), X. Bailly (2)

(1) Laboratoire de Physique Corpusculaire, CNRS/IN2P3, 63000 Clermont Ferrand, France
(2) INRA Clermont-Theix, 63122 SAINT-GENES CHAMPANELLE, France

Introduction :

Rhizobium leguminosarum est une espèce bactérienne capable de fixer l'azote atmosphérique au bénéfice de ses plantes hôtes [1]. Au sein de cette espèce, différentes souches sont capables de s'associer avec des espèces hôtes différentes de Légumineuses, et cette étude se concentrera sur la vesce (*Vicia sativa*) et le trèfle (*Trifolium repens*). Le but de cette étude est de caractériser la différenciation génétique entre des souches de *Rhizobium leguminosarum* associées au trèfle ou à la vesce par une approche pangénomique. Plus précisément, la mesure de différenciation calculée évalue la distance génétique moyenne entre souches issues de plantes hôtes différentes par rapport à la distance génétiques moyenne entre deux souches de *R. leguminosarum*. L'objectif de ce travail est de déterminer si cette différenciation est hétérogène le long du génome et, à terme, d'identifier des gènes dits « candidats» où la différenciation est particulièrement forte, qui pourraient être liés aux mécanismes d'adaptation de la bactérie vis-à-vis des plantes hôtes.

Dans ce but, nous avons créé une chaîne de traitements permettant, à partir d'éléments issus de l'assemblage *de-novo* des séquences provenant de 36 souches de *Rhizobium leguminosarum bv trifolii* et de 36 souches de *Rhizobium leguminosarum bv viciae* :

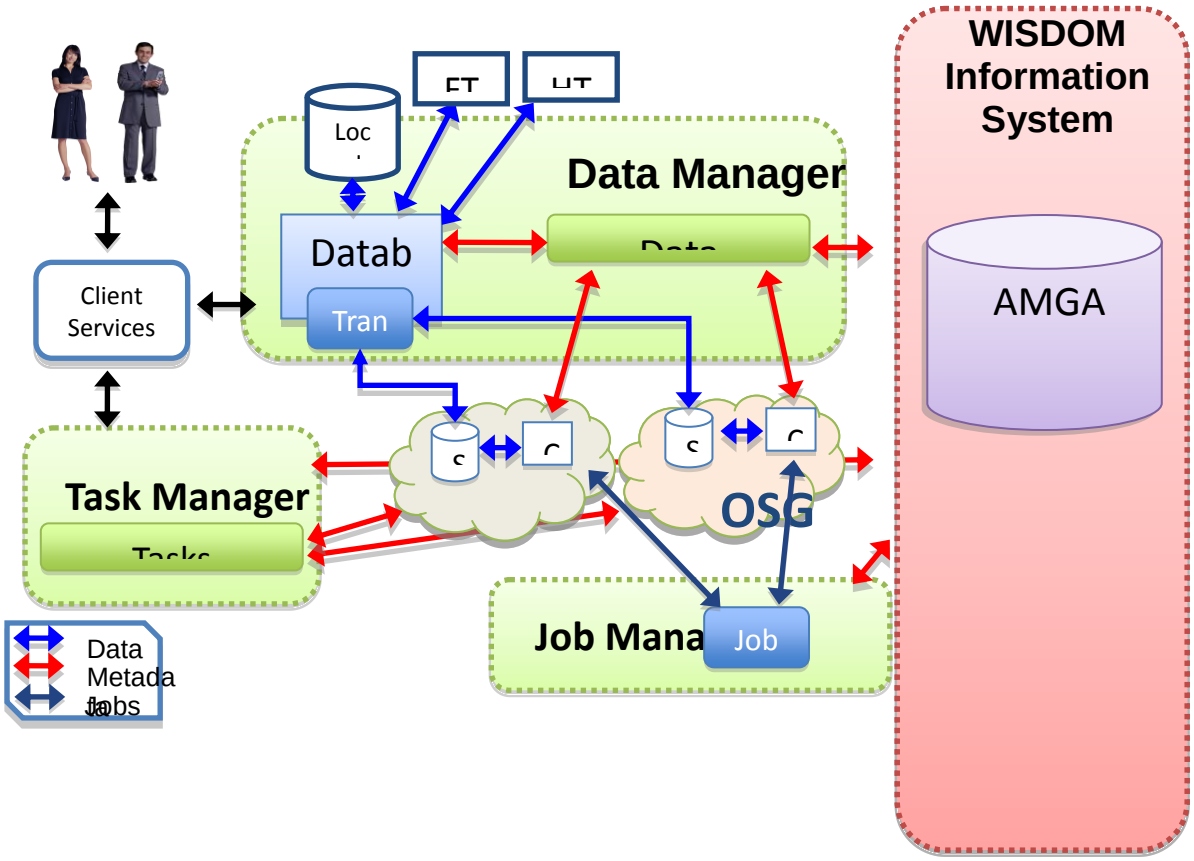
- d'extraire les séquences des souches étudiées pour les différents gènes annotés sur les génomes de *Rhizobium leguminosarum* disponibles dans GenBank
- de réaliser un alignement multiple des séquences extraites pour chaque gène
- d'obtenir un arbre phylogénique pour chaque alignement
- de calculer pour chaque gène le niveau de différenciation

Ces différentes analyses demandent un temps d'exécution important, notamment la construction des arbres phylogénétiques. En effet, il est nécessaire de créer autant d'arbres que de gènes, à savoir plus de 10 000. De plus, il faut ajouter tous les fichiers intermédiaires utilisés lors du traitement. A chaque étape du processus, le résultat doit être analysé afin d'évaluer la pertinence de la séquence par rapport à ce qui était recherché.

Cela correspond environ à 75 000 fichiers pour l'intégralité de notre recherche. Il est impossible de lancer une telle simulation sur un cluster de calcul, et c'est pourquoi nous nous sommes orientés vers la grille.

Pour pouvoir tirer le meilleur partit des ressources de la grille, l'environnement de production WISDOM [2] a été utilisé. Cet environnement en « pull model » nous a permis d'optimiser les performances vis-à-vis de la soumission. Etant donné que chaque étape du pipeline nous donnait des résultats nécessaires pour la suite, et vu que les temps de calcul évoluaient en fonction des séquences que nous avions en entrée, ce système de production a permis une meilleure progression dans nos résultats ainsi qu'une répartition de la charge de travail accrue.

L'environnement de production WISDOM



L'environnement se découpe en 3 parties :

- Le job manager : responsable de la soumission des jobs, il a pour rôle de s'assurer d'un nombre de ressources suffisant pour traiter l'ensemble des tâches qui sont soumises. Plus il y a de tâches en attente, plus le nombre de jobs doit augmenter. Une fois que ce nombre décroît, le nombre de jobs (agents) décroît en parallèle.
- Le data manager : responsable de la copie, de la réplication et de la récupération de toutes les données envoyées sur la grille, notamment tout ce qui concerne les bases de données biologiques. Il est découpé en 2 parties, une utilisable par les utilisateurs (copie, récupération), une automatique (réplication des bases de données sur les différents SE)
- Le task manager : responsable du traitement des tâches (calculs) soumis par les utilisateurs de l'environnement (par exemple 1000 calculs de Blast). Chaque tâche est basée sur 2 éléments : un service (correspondant à un package comprenant le programme à lancer), et des données en entrée (copiées sur les SE grâce au Data Manager)

Le flux des données :

Ce flux se décompose en sept grandes parties :

- Une sélection des gènes à utiliser parmi les génomes des espèces étudiées. Cette étape permet d'éviter la redondance (c'est-à-dire évite qu'on fasse une analyse d'un même gène présents pour différentes espèces).
- Une étape de comparaison des fractions de génomes des 72 espèces bactériennes sur les séquences homologues des gènes retenus en utilisant l'algorithme BLAST. Cela permet de retrouver à quel gène appartenait chaque contig
- Un assemblage des différentes fractions étudiées pour reconstituer les séquences des gènes d'origine à l'aide de programmes spécifiques en langage Perl .
- Regroupement de toutes les séquences d'un même gène chez les différentes espèces dans un même fichier
- Un alignement global de toutes les séquences pour chaque gène
- La création pour chaque gène d'un arbre phylogénétique
- Le calcul du Kst permettant de déterminer si les espèces bactériennes se regroupent en fonction de leur plante hôte d'origine ou non.

Ces différentes parties sont regroupées en 5 étapes :

- Fusion des listes de gènes : exécution en local
- Comparaisons et assemblage : exécution sur grille, service WISDOM BlastAndSort
- Programme spécifique DispatchGenes : exécution en local
- Alignement global et création de l'arbre phylogénique : exécution sur grille, service WISDOM MusclePhyml
- Calcul des Kst : exécution en local, AnalyseDiffTree.pl (Programme spécifique en langage Perl)

Dans ce flux, , la Grille a été utilisée de la manière suivante :

- Création des arbres phylogénétiques
- Création de scripts (services) WISDOM pour l'assemblage
    - Analyse des correspondances entre les séquences
    - Reconstruction de séquences consensus

➢ Fusion des listes de gènes :

Cette fusion consiste en trois BLAST successifs des listes de gènes les unes contre les autres. Cela permet d'identifier si les gènes sont présents dans les différents génomes et de n'en garder qu'une seule copie. Elle permet aussi d'éliminer les gènes homologues et donc très proches.

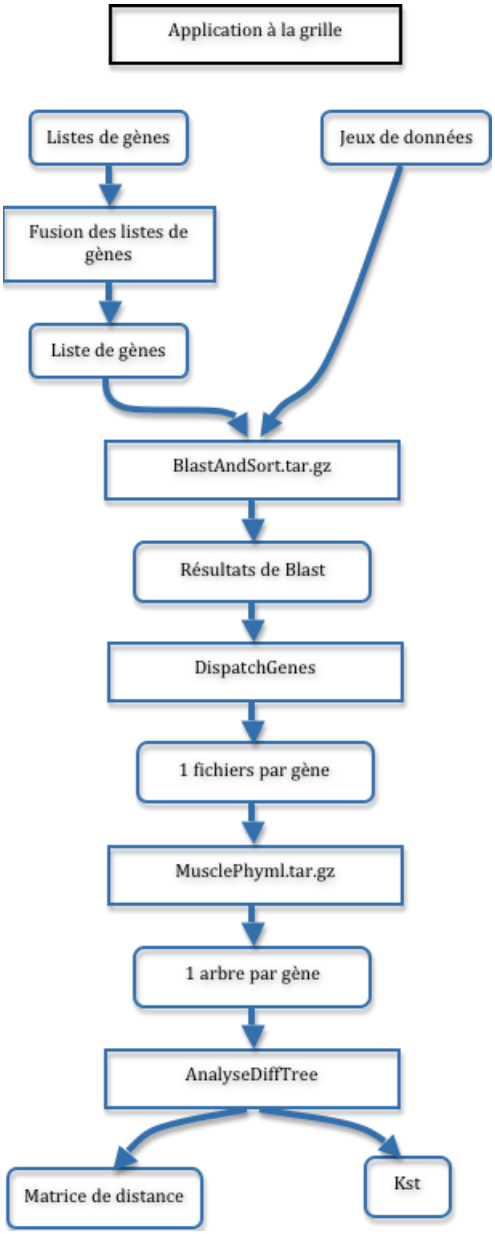
Résultats : grâce à cette étape, le nombre de gènes à analyser est passé de 20559 à 10368, ce qui constitue une diminution de la quantité de données à analyser de près de 50%.

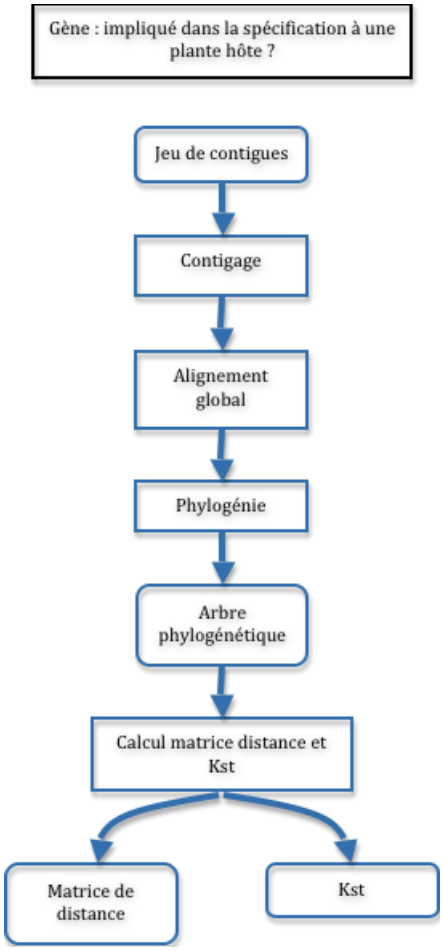
➢ BlastAndSort:

Tout d'abord le service effectue un BLAST du fichier de contig d'une espèce sur la liste de gènes précédemment créée. Par la suite il effectue plusieurs tris des résultats de BLAST pour les associer en fonction de la du pourcentage d'identité obtenu et sa significativité. Ensuite, il analyse

les contigs retenus pour chaque gène et sélectionne les meilleurs candidats pour être assemblés. Une fois tout les contigs intéressants sélectionnés, le programme les assemble les uns aux autres.

Résultats : Le service a été utilisé sur les 72 fichiers de contigs appartenant chacun à une espèce. Chaque analyse génère 9 fichiers, soit au total 648 fichiers pour une taille total de 417 Mo. Grâce à la grille, cette analyse, dont le temps était initialement estimée à 24h, a été exécutée en environ 1h30.





- **DispatchGenes :**

Ce programme prend en entrée les fichiers contenant les contigs assemblés et les trie en fonction du gène d'origine auquel ils ont été associés. De cette façon, il y a un fichier pour chaque gène ayant des contigs assemblés.

Résultats : Suite à cette étape, le script a généré 8486 fichiers de tailles variables. Cela signifie que le nombre de gènes étudiés passe de 10368 à 8486.

- **MusclePhyml:**

Ce module réalise l'alignement et la phylogénie. Pour commencer, elle effectue un alignement global des séquences précédemment obtenues sur chaque gène mais aussi de la séquence originale du gène. En effet, ceci permet au programme d'alignement global, Muscle, de donner des résultats cohérents car les séquences reconstituées comportent parfois de longues insertions/délétions qui peuvent rendre les résultats d'alignement peu fiables.

Une fois les fichiers de sortie obtenus, ceux-ci sont remaniés pour être adaptés (passage au format PhyLip) au programme Phyml qui permet de construire les arbres phylogénétiques.

Résultats : L'analyse précédente ayant généré de nombreux résultats, il a fallu les regrouper pour former 84 archives contenant les 8486 analyses à faire. L'analyse de gènes a créé 5 fichiers. Le programme à généré au total 42430 fichiers en une nuit au lieu de 20 jours.

- **AnalyseDiffTree.pl**

L'étude de la différentiation au sein des arbres de phylogénique a été réalisée en calculant le Kst . Pour cela le programme calcule en premier lieu la matrice de distance de l'arbre. A la fin de ce travail, 75368 fichiers ont été générés

Conclusion :

Lors de cette analyse certains gènes d'intérêt, gènes ayant un fort Kst (cf AnalyseDiffTree) ont pu être mis en évidence. Cependant, un biais à été déterminé après l'analyse de résultats. En effet, certains gènes se retrouvaient uniquement chez une seule espèce bactérienne. Cette situation est produite à l'issue du traitement de comparaison avec le BLAST lors de la sélection des gènes car certains d'entre eux sont homologues. Une étude plus approfondie de cette étape est donc nécessaire. De même, la production de plusieurs milliers d'arbres sur plusieurs milliers de gènes fournit un matériel qui reste grandement exploiter. Ceci fera l'objet de travaux plus poussés au sein de l'Unité dans le futur.

Bibliographie

1. Xavier Bailly, Isabelle Oliviery, Brigitte Brunel, et al. 2007. Horizontal Gene Transfert and Homologous Recombination Drive The Evolution of the Nitrogen-Fixing Symbionts of Medicago Species. J. Bacteriol. 189(14) :5223-5236
2. Jacq, N., Salzemann, J., Jacq, F., Legrés, Y., Medernach, E., Montagnat, J., Maa, A., Reichstadt, M., Schwichtenberg, H., Sridhar, M., Kasam, V., Zimmermann, M., Hofmann, M., Breton, V., Grid-enabled Virtual Screening against malaria, to be published in Journal of Grid Computing, (2007).